複数短単位版「分類語彙表番号-UniDic」対応表の整備と公開

片山 久留美 [1]　髙橋 雄太 [12]　菊池 そのみ [3]　小木曽 智信 [1]
[1] 人間文化研究機構 国立国語研究所　[2] 明治大学　[3] 筑波大学
{kurumi_katayama,ytaka,togiso}@ninjal.ac.jp　kikuchi.sonomi.gn@u.tsukuba.ac.jp

概要

コーパスに語義を付与することのできる語彙資源として「分類語彙表番号-UniDic 語彙素番号対応表」が公開されているが，収録語は分類語彙表の見出し語が UniDic 短単位と完全一致するものに留まっていた．本研究では，これを補完するデータとして複数短単位に分割される分類語彙表見出しと UniDic との対応表を作成した．このデータを活用することで，国立国語研究所のコーパスや短単位解析結果に対して本来の分類語彙表番号を付与することが可能になったほか，分類語彙表見出し語の構造について UniDic の情報を利用した分析を行うことが可能となった．本データの構築方法や公開形式について報告する．

1　はじめに

近年，言語研究・自然言語処理の分野において日本語のコーパスに語義情報を付与することが求められている．国立国語研究所では，大規模な現代日本語のシソーラスである国立国語研究所（2004）『分類語彙表増補改訂版データベース』（ver.1.0）[1]（以下「分類語彙表 DB」と呼ぶ）の見出し語と，国立国語研究所で公開するコーパスの形態素解析に用いる形態素解析辞書「UniDic」の見出し語との同語関係に基づく対応表「分類語彙表番号-UniDic 語彙素番号対応表」[2]（以下「1 短単位版対応表」と呼ぶ）を公開した．しかし，「1 短単位版対応表」は「分類語彙表 DB」の見出し語と UniDic の見出し語の単位である「短単位」とが一致したもののみを対象としたものであった．そこで，「分類語彙表 DB」の見出し語のうち，「1 短単位版対応表」には未収録となった複数の短単位から成る見出し語に対する UniDic 語彙素番号との対応表を新たに作成した．これにより，「分類語彙表 DB」の全ての見出し語と UniDic の語彙素との対応表を構築することができた．本発表では，複数の短単位から成る分類語彙表見出し語と UniDic との対応付けの手法と，データの概要について報告する．

2　「分類語彙表番号-UniDic 語彙素番号対応表」の概要

「分類語彙表 DB」では，各見出しに対して「分類番号」を付与している．「分類番号」は「1.3131」のような 5 桁の数字として表記される．「類」「部門」「中項目」「分類項目」という 4 階層の意味的範疇を示す構造となっており，さらに分類番号と分類項目の中を分類する「段落番号」「小段落番号」，および「小段落番号」内の配列順序を表す「語番号」に基づいて見出しを配列している [3]．「分類番号」「段落番号」「小段落番号」「語番号」の 4 つが揃うことで見出しを一意に指定することができる．「1 短単位版対応表」ではこの 4 つの番号を連結した「分類語彙表番号」を各見出しの ID として付与している．

一方，「UniDic」は国立国語研究所が整備している電子化辞書である [4]．UniDic では，「短単位」に基づいて語を認定・登録している．「短単位」は，現代語において意味を持つ最小の単位（「最小単位」）を認定したうえで，「最小単位」の 0 回または 1 回の結合によって得られる単位である．言語の形態論的側面に着目した揺れの少ない斉一な単位 [5] であり，コーパスにおける用例収集に適している．また，UniDic の各見出し語は，表記や語形の違いにかかわらず同語と考えられる語については同一の見出しを与えるという方針に基づいて語の登録を行っており，階層的な構造を持っていることに特徴がある．最上層として国語辞典の見出しに相当する「語彙素」を置き，その下の階層に語形のゆれを区別する「語形」，さらにその下層に異表記を区別する「書字形」を設けている．たとえば，図 1 に示すように，語彙素「攪拌（カクハン）」には「カクハン」「コウハン」という 2 つの語形があり，さらに各語形の下に実際の表記形である「書字形」および「発音形」が登録されている．こうした語形や書字形などのゆれを 1 つにまとめ上げるのが語彙素であり，各語彙素には固有の ID 番号である「語彙素 ID」が付与されている．

語彙素ID	語彙素 （語彙素読み）	語形	書字形	発音形
42895	攪拌 （カクハン）	カクハン	攪拌	カクハン
			撹拌	
			かく拌	
		コウハン	攪拌	コーハン

図 1　UniDic の階層構造の例

「1 短単位版対応表」は,「分類語彙表 DB」と「UniDic」の見出し語のうち，同語関係にあると考えられるものの「分類語彙表番号」と「語彙素 ID」を対応づけた表である．その構築にあたっては,「分類語彙表 DB」全 98,241 見出しのうち，1 短単位から成ると見られる見出し語を人手により選別して作られている [6]．1 短単位から成る見出し語とは，たとえば「走る」「元気」などのように「分類語彙表 DB」の見出しが UniDic に登録されている短単位の切れ目と一致するものを指す．「1 短単位版対応表」によって,「分類語彙表 DB」全 98,241 見出しのうち約 66 %に当たる 64,759 見出しについて「UniDic」の語彙素 ID との対応を取ることが可能となっている．

「1 短単位版対応表」の対象外となった見出しは,「協力する」「異端児」など短単位認定規定上 1 短単位と認められない語や,「なくてもいい」「なるべくなら」のような連語・複合辞，また「痛くもない腹を探られる」のような慣用表現が含まれている．これら複数短単位から成ると見られる 33,477 見出しおよび短単位未満の単位から成る 5 見出しについても UniDic の見出し語との対応付けを行うことで,「分類語彙表 DB」全見出しについての「対応表」を作成することが本研究の目的である．

3　複数短単位から成る見出し語に対する対応表の構築

3.1　見出し語の形態素解析

複数短単位版の対応表を作成するためには，まず「分類語彙表 DB」の見出し語を短単位に分割する必要がある．そこで,「現代書き言葉 UniDic」を用いた形態素解析を行うことで見出しを短単位に分割し，UniDic の情報を付与することとした．対象となる 33,482 見出しを XML 化し，タグの属性値に分類語彙表番号を付与したうえで形態素解析を行い，得られた形態論情報について国立国語研究所の「形態論情報データベース」[7] 上で人手による修正を行った．

3.2　活用語・異語形に対する対応

複数短単位から成る見出しの特徴として，活用語が終止形（基本形）以外の形式で現れることが挙げられる．たとえば，形容詞「良い」は,「威勢がよい」という見出しの場合は文末での終止用法が想定される．しかし,「頭のよい」という見出しの場合，連体修飾用法で使用されると考えられる．1 短単位から成る見出しの場合には，基本的に辞書の見出し語のように言い切りの形で現れるものが多かったため活用形については問題にならなかった．しかし，複数短単位版ではこうした活用形の情報が重要となる．

また,「分類語彙表 DB」の見出しでは「攪拌（カクハン）」に対する「攪拌（コウハン）」のように，UniDic の語彙素レベルでは同語と判定されるものの異語形が別見出しとして現れることがある．今回のデータ構築に当たっては，UniDic による形態素解析を行ったという利点を生かし，こうした活用形や語形レベルの違いについても形態論情報に反映させることとした．これにより，後述する語彙表 ID（lid）を活用した詳細な分析に対応したデータを作成することが可能となった．

- 「威勢がよい」
  威勢　語彙素「威勢」名詞-普通名詞-一般
  が　語彙素「が」助詞-格助詞
  よい　語彙素「良い」形容詞-非自立可能　形容詞　終止形-一般
- 「頭のよい」
  頭　語彙素「頭」名詞-普通名詞-一般
  の　語彙素「の」助詞-格助詞
  よい　語彙素「良い」形容詞-非自立可能　形容詞　連体形-一般
- 「攪拌（カクハン）する」
  攪拌　語彙素「攪拌」 語形「カクハン」 発音形「カクハン」
  する　語彙素「為る」
- 「攪拌（コウハン）する」
  攪拌　語彙素「攪拌」 語形「コウハン」 発音形「コーハン」
  する　語彙素「為る」

4 「複数短単位版対応表」の概要

4.1 語彙統計

「複数短単位版対応表」33,482 見出しについて，「分類番号」の最上層である「類」ごとに分類すると，表 1 のようになる。体の類が全体の約 46 %, 用の類が約 43 %を占めている．

表 1 「類」別の見出し語数

類	見出し語数
1 体の類	15310
2 用の類	14280
3 相の類	3466
4 その他の類	426
総計	33482

また，1 つの「分類語彙表 DB」見出しがいくつの短単位に分割されたかを集計したのが表 2 である[1)]．2 短単位に分割された見出しが 25,743 件と全体の約 77 %を占める．最大では 9 短単位に分割された見出しがある．9 短単位となったのは，「居ても立っても居られない」「一つかまの飯を食った仲」「目の中に入れても痛くない」「横の物を縦にもしない」の 4 見出しであった．

2 短単位に分割された 25,743 件のうち，9,603 件は用の類で語彙素「為る」を含むサ行変格活用の複合動詞である．「勉強する」「アウトプットする」「いらいらする」など，様々なサ変複動詞が登録された．

表 2 構成短単位数別の見出し語数

構成短単位数	見出し語数
1	109
2	25743
3	6162
4	989
5	364
6	78
7	27
8	6
9	4
総計	33482

1) 構成短単位数 1 となっているのは，「1 短単位版対応表」作成時には複数短単位になると見られ対象外とされたものの，今回のデータ整備において 1 短単位と認定されたものである．

4.2 分類語彙表と UniDic の多対多対応

「分類語彙表 DB」の見出しを複数短単位に分割し形態論情報を付与すると，「分類語彙表 DB」では別の分類番号を与えられ別見出しになっていても，UniDic においては同一の語彙素 ID の組み合わせから成ると判断されるものが見られる．つまり分類語彙表番号と UniDic の語彙素 ID が多対一で対応するものである．

たとえば，「平行する」という見出しは「分類語彙表 DB」において 5 つの分類番号を付与されている．しかし，UniDic の語彙素ではいずれも語彙素「平行」と語彙素「為る」の組み合わせとして処理される．複数短単位版対応表では，3727 パターンの語彙素 ID の組み合わせにおいて複数の分類番号との対応が見られた．

「平行」(語彙素 ID33855) ＋「為る」(語彙素 ID19537)
2.1120 関係-類-相対
2.1525 関係-作用-連れ・導き・追い・逃げなど
2.1570 関係-作用-成形・変形
2.1573 関係-作用-配列・排列
2.1730 関係-空間-方向・方角

一方，1 つの分類語彙表見出しに対して複数の UniDic 語彙素が対応する一対多の対応パターンは，「複数短単位版対応表」ではほとんど見られなかった．「1 短単位版対応表」においても多対一の対応が一対多の対応より多く見られることが指摘されているが，「複数短単位版対応表」では一対多の対応例がさらに現れにくいことが指摘できる[2)]．

5 「複数短単位版対応表」の利点

「複数短単位版対応表」を活用することによって，より正確に「分類語彙表 DB」の意味分類を使用した分析が可能となる．たとえば，図 2 に示す作例「頭のよい大学生が一堂に会し議論する」は 11 の短単位から成る文である．「一堂」は「分類語彙表 DB」に単独の語としての記載がないために，「1 短単位版対応表」に掲載されていない．しかし，「複数短単位版対応表」では「一堂に会する」で 1 見出しとなっているため，適切な分類語彙表番号を付与することが可能となる．

2) 「1 短単位版対応表」で一対多対応より多対一対応が多く見られることについては，近藤・田中（2020）が「分類語彙表 DB が多義語の意味ごとに見出しを立て日本語の表しうる意味の世界を示そうとするシソーラスであるのに対し，UniDic が微妙な意味の差やそれに対応する語の書き分けをできるだけまとめあげて 1 語彙素とし，形態素解析の精度を保持しようとする形態素解析辞書用データであるという両者の使用目的の違いによりもたらされたもの [6]」と指摘している．

複数短単位版分類番号	3.3421			1.2419			2.3510			2.3133	
短単位	頭	の	よい	大学	生	が	一堂	に	会し	議論	する
語彙素ID	741	28989	38988	22965	19690	7889	2096	28178	5775	9967	19537
1短単位版分類番号	1.3421		3.1332	1.2630					2.3510	1.3133	2.3320

図 2　複数短単位版対応表と 1 短単位版対応表の分類番号

また，「議論する」のような漢語サ変動詞については，「1 短単位版対応表」を用いた分析では「議論」「する」の 2 語として扱わざるを得なかった．「1 短単位版対応表」における動詞「する」は，対応する分類語彙表番号が 8 つある多義語である．このため，文脈等を踏まえないと語義を一つに絞ることができないという問題があった．しかし，4.1 節で述べたように，「複数短単位版対応表」では多くのサ変複合動詞が新たに登録された．これを用いることで，より正確に対応する分類語彙表番号を取得することが可能となった．

6　データの公開形式

「複数短単位版対応表」は用途に応じて活用できるよう，複数の公開形式を用意している．まず，「1 短単位版対応表」と同様の形式として，分類語彙表番号と UniDic の語彙素 ID を並べたテキストデータ形式である．語彙素 ID を UniDic の辞書アーカイブ（現代語用 UniDic のフルパッケージ）[3] の語彙素一覧（lex.csv）に含まれる語彙素 ID 列に対応付けることで，語彙素・語彙素読みや語種，品詞の上位概念を示す類などの情報を取り出すことが可能である．

分類語彙表見出し, 分類語彙表番号, 短単位数, 語彙素 ID-1, 語彙素 ID-2, 語彙素 ID-3

威勢がよい,3.3430-02-02-02,3,1949,7889,38988

「短単位数」とは，分類語彙表見出しがいくつの短単位に分割されたかを示す数値である。上記の例では 3 短単位に分割されていることを表す．「1 短単位版対応表」と同様に語彙素レベルでの語の同定を目的とする場合には，このデータが有用である．

次に，分類語彙表番号と語彙表 ID を並べたテキストデータ形式の対応表について述べる．3.1 節で述べたように，「複数短単位版対応表」では語形の違いや活用語の活用形まで区別して形態論情報の付与を行っている．語彙素 ID はあくまで語彙素を一意に識別するものであるため，こうした語形や活用形についての情報は語彙素 ID から得ることができない．そこで，分類語彙表番号に対して UniDic の語彙表 ID（lid）を対応させたテキストデータも公開することとした．語彙表 ID（lid）とは，「UniDic DB 中の各エントリ（短単位）を一意識別するための ID[4]」[7] である．語彙素 ID とは異なり，語形・書字形・発音形・活用形などの違いも区別することができる．「威勢がよい」「頭のよい」の例では，語彙素 ID による対応表では「よい」にはいずれも「38988」という同じ ID が付与されるが，語彙表 ID（lid）では活用形を表す末尾部分の数字が異なる．語彙表 ID を活用することで，より詳細に語の情報を得ることが可能となる．

さらに，テキストデータだけではなく XML 形式のデータも公開する．XML 形式のデータでは，分類語彙表の各見出し語を単位としてそれを構成する短単位を子要素とし，短単位の形態論情報を属性値として記述している．語彙素 ID，語彙素，語形，品詞，発音形，語種，語彙表 ID を他のデータと関連付けることなく参照することができる．

これらのデータは，2023 年度内に国立国語研究所学術情報リポジトリにて公開予定である．

7　おわりに

「複数短単位版対応表」を作成したことで，「分類語彙表 DB」の全見出しに対する UniDic の語彙素 ID との対応付けが実現した．語彙表 ID による対応表を新たに作成したことで，語形や活用形といった UniDic が持つ語に関する詳細な情報が活用できる．「1 短単位版対応表」と合わせて，言語研究・自然言語処理等の分野等で活用することが期待される．

3）　https://clrd.ninjal.ac.jp/unidic/back_number.html

4）　UniDic 用語集: 語彙表 ID と語彙素 ID https://clrd.ninjal.ac.jp/unidic/glossary.html#lid_and_lemma_id

謝辞

本研究は，国立国語研究所共同研究プロジェクト「多様な語彙資源を統合した研究活用基盤の共創」の成果の一部です．

参考文献

[1] 国立国語研究所. 分類語彙表増補改訂版データベース（ver.1.0.1）, 2018. https://github.com/masayu-a/WLSP.

[2] 国立国語研究所. 分類語彙表番号 − unidic 語彙素番号対応表, 2020. https://github.com/masayu-a/wlsp2unidic.

[3] 国立国語研究所. 分類語彙表増補改訂版. 大日本図書, 2004.

[4] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. 日本語科学, No. 22, pp. 101–123, 2007.

[5] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上）（下）. 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書, 2011.

[6] 近藤明日子, 田中牧郎. 「分類語彙表番号 − unidic 語彙素番号対応表」の構築. 国立国語研究所論集,No.18, pp.77-91, 2020.

[7] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用. 自然言語処理, Vol. 21, No. 2, pp. 301–332, 2014.

A　語彙素 ID (lemmaID) テキスト

分類語彙表見出し, 分類語彙表番号, 短単位数, 語彙素 ID-1, 語彙素 ID-2, 語彙素 ID-3
威勢がよい,3.3430-02-02-02,3,1949,7889,38988
頭のよい,3.3421-03-02-01,3,741,28989,38988

B　語彙表 ID (lid) テキスト

分類語彙表見出し, 分類語彙表番号, 短単位数, 語彙表 ID-1, 語彙表 ID-2, 語彙表 ID-3
威勢がよい, 3.3430-02-02-02, 3, 535745664262656, 2168520431510016, 10716957049496235
頭のよい, 3.3421-03-02-01, 3, 203693219783168, 7968444268028416, 10716957049496257

C　XML

```
<wlspWord headword="威勢がよい" code="3.3430-02-02-02" units="3">
    <SUW lemmaID="1949" order="1" lemma="威勢" lForm="イセイ" pos="名詞-普通名詞-一般" pron="イセー" wType="漢" lid="535745664262656" />
     <SUW lemmaID="7889" order="2" lemma="が" lForm="ガ" pos="助詞-格助詞" pron="ガ" wType="和" lid="2168520431510016" />
    <SUW lemmaID="38988" order="3" lemma="良い" lForm="ヨイ" pos="形容詞-非自立可能" cType="形容詞" cForm="終止形-一般" pron="ヨイ" wType="和" lid="10716957049496235" />
</wlspWord>
<wlspWord headword="頭のよい" code="3.3421-03-02-01" units="3">
    <SUW lemmaID="741" order="1" lemma="頭" lForm="アタマ" pos="名詞-普通名詞-一般" pron="アタマ" wType="和" lid="203693219783168" />
     <SUW lemmaID="28989" order="2" lemma="の" lForm="ノ" pos="助詞-格助詞" pron="ノ" wType="和" lid="7968444268028416" />
    <SUW lemmaID="38988" order="3" lemma="良い" lForm="ヨイ" pos="形容詞-非自立可能" cType="形容詞" cForm="連体形-一般" pron="ヨイ" wType="和" lid="10716957049496257" />
</wlspWord>
```